Journal of Medicinal Plants Studies
www.PlantsJournal.com

ISSN (E): 2320-3862
ISSN (P): 2394-0530
NAAS Rating 2017: 3.53
JMPS 2017; 5(2): 143-147

Received: 26-01-2017
Accepted: 27-02-2017

Anirban Goswami
Investigator (Statistics), Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India

Dr. Mohd Wasim Ahmed
Research Officer (U), Scientist L-1, Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India

Dr. Rajesh
Research Officer (U), Scientist L-1, Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India

Dr. Mohd Ishtiyaque Alam
Research Officer Incharge, Scientist L-4, Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India

Dr. Hashmat Imam
Senior Research Fellow (U), Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India

Dr. Aisha Perveen
Senior Research Fellow (U), Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India

Correspondence
Anirban Goswami
Investigator (Statistics), Regional Research Institute of Unani Medicine, Patna, under CCRUM, Ministry of Ayush, India

Disease pattern discover in institutional data via cluster analysis

Anirban Goswami, Dr. Mohd Wasim Ahmed, Dr. Rajesh, Dr. Mohd Ishtiyaque Alam, Dr. Hashmat Imam and Dr. Aisha Perveen

Abstract
In this presents study to discover the disease patterns by using statistical approach on institutional database. This study, help to know about the number of patients attended in OPD with different ailments for every year. By the help of data mining technique, the study is designed to discover the patterns and hidden relationships in dataset. Actually data mining technique is used to extract information from a data set and transform it into an understandable structure for further uses. The main aims of this study to provide profiling of patients, discover dominant disease and dominant month via cluster analysis. A cluster is a collection of data objects which are similar to one another within the same cluster and are dissimilar to the objects in other clusters. In this regard, clustering is used to profile patients according to their month of GOPD in the institute. The gap statistic used to find the optimum numbers of clusters in dataset. Using this, a number of clusters are formed on the basis of type of disease acquired by patients, demographic and socioeconomic characteristics beside that patients are grouped into different clusters according to their disease.

Keywords: Disease pattern, Clustering, Segments, Dominant, Optimum

Introduction
Data mining is an important analytic process designed to explore data. It is a process of analyzing from different perspectives and which finds useful patterns from large amount of data. Data mining is a process of extraction of useful information and patterns from huge data. Data mining is the core part of the knowledge discovery in database [1]. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining technique lies at the interface of statistics, database technology, pattern recognition, machine learning, data visualization, and expert systems. A database is a collection of data that is organized/institute so that its contents can easily be accessed, managed, and updated. Databases contain aggregations of data records or files, and a database manager provides users the capabilities of controlling read and write access, specifying report generation, and analyzing use. The actual data mining is the process of automatic or semiautomatic analysis of large set of data to retrieve previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining).
Particular attention will be paid to the results of cluster analysis on a subset of data from one of the larger institutes, where several patient subgroups have been tentatively identified. Cluster analysis is a technique of choice to search for unknown groups in a population. However, cluster analysis can be a challenging multivariate approach when clusters are marginally separated. Clustering method is to form the clusters from large database on the basis of similarity measure [2]. The goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic [2]. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belongs to different groups. Clustering approach is used to identify similarities between data points. Each data points within the same cluster are having greater similarity as compare to the data points belongs to other cluster. Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. There are many clustering algorithm available today to gather the

data by comparing the similarities between the data and analyzing their distance matrix etc. The algorithm in use is hierarchical, agglomerative, ward's method and K-means etc.
Today most of peoples suffered from various types of diseases. Sometimes especially the topical or seasonal diseases become viral and every second person found ill. Lot of patients lost their health condition due to lack of treatment. The facilities of institute collapsed very rapidly because number of patients suddenly increases and institute have arrangements to provide treatment or medicine to patients. In this study mainly focuses on these types of problems faced by both patients and institute administration. This study finds out the dominant disease i.e. by which disease optimum number of dominant patients suffered and in which dominant month so that institute management can increase or decrease its facilities like number of beds, medicines, doctors, specialists etc. In time and all the patients can get the treatment easily and no one have to lost their health condition. Data Mining technologies and Clustering are to provide benefits to healthcare organization/ institute for grouping the patients having similar type of diseases or health issues so that healthcare organization /institute provides them effective treatments. Data Mining and Clustering are also used to analyze the various factors that are responsible for diseases by demographic factors or socioeconomic factors.

2. Methodology

The all medical records including demographic and socioeconomic data (such as age, sex, occupation…etc) of patients are store in OPD management software at Regional Research Institute of Unani Medicine, Patna during year 2015-16. The use of 'R' software for access the SQL database from OPD management software [3]. The 'R'(version 3.3.2) and "WEKA" (version 3.6) software use to making the profiling of patients,discover dominant disease and dominant month. In this study k-means cluster analysis used to search for unknown groups in a patients and to profiling of patients using patient's demographic and socioeconomic data.
A major challenge in cluster analysis is the estimation of the optimum number of cluster in dataset. In this regard the gap statistic use to determine the optimum cluster [4, 5]. The idea behind their approach was to find a way to standardize the comparison of $\log W_k$ with a null reference distribution of the data, i.e. a distribution with no obvious clustering. Estimate for the optimal number of clusters K is the value for which $\log W_k$ falls the farthest below this reference curve. This information is contained in the following formula for the gap statistic: $Gap_n(k) = E_n^*\{\log W_k\} - \log W_k$. To obtain the estimate $E_n^*\{\log W_k\}$, first compute the average of B copies $\log W_k^*$ for $B=50$, each of which is generated with a Monte Carlo sample from the reference distribution. Those $\log W_k^*$ from the B Monte Carlo replicates exhibit a standard deviation sd_k which, accounting for the simulation error, is turned into the quantity $s_k = \sqrt{(1+\frac{1}{B})}\, sd_k$.Finally, the optimal number ofclusters K is the smallest k such that $\text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1}$.
The computation of the gap statistic involves the following steps Cluster the observed data, varying the number of clusters from $k = 1,\ldots, k_{max}$ and compute the corresponding W_k.

- Generate B reference data sets and cluster each of them with varying number of clusters $k = 1,\ldots,k_{max}$. Compute the estimated gap statistic:
  $\text{Gap}(k) = \left(\frac{1}{B}\right)\sum_{b=1}^{B} \log W_{kb}^* - \log W_k$; $k=1,2,\ldots,K$..
- With $\bar{w} = \left(\frac{1}{B}\right)\sum_b \log W_{kb}^*$ compute the standard deviation $sd_k = [\left(\frac{1}{B}\right)\sum_b \{\log W_{kb}^* - \bar{w}\}^2]^{1/2}$ and define $s_k = \sqrt{(1+\frac{1}{B})}\, sd_k$.
- Choose the number of clusters as the smallest k such that $\text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1}$.

After estimation of the optimum cluster used to K–means clustering to grouping the institutional data. The algorithm and it has the following general form [6, 7]:

(i) Select estimated 'k' cluster centers by using gap statistic.
(ii) Calculate the distance between each data point and cluster centers.
(iii) Add each element of the remaining items to the part that it's the cluster center is minimum of all the cluster centers, and used Euclidean distance to measure the dimension.
(iv) Select new centers of parts that have been formed.
(v) Regroup items about the new centers, and then define the new centers.
(vi) Repeats the above steps till settle the distribution, which don't change centers.

Discriminate analysis used after k-means cluster analysis to classify cluster. In the visualization of cluster classification, the cluster observations are represented by points to plot in the graph, using principal components. An ellipse is drawn around each cluster in the graph.
The main problems related to disease pattern discover in institutional data. These problems are following as:

2.1 Profiling of patients

In this particular problem, the patients in institute are segregated into different groups to get information about the number of patients attended in GOPD. Then patients are segregated into different groups according to their month segment of attended in GOPD. Consider, M_1, M_2,…,M_n are various clusters refers different segments of month. Here month segment means number of patients attended in every three months. From its result, the dominant month and the dominant month segment are known to us i.e. in which month and in which month segment optimum numbers of patients were attended.

2.2 Disease Pattern Discover via Clustering

Dataset of different type of disease segments (which are found by profiling), are collect and feed to Weka software to find out the dominant month in which number of patients of particular disease is optimum. From here a link between month and disease segments can be found. In addition, in a month there are different type of patients are attended but we have to find out the dominant disease i.e. the name of disease having optimum patients in that particular month/month segment. The same database is feed to Weka software and form four clusters. These clusters contain information about month segments as well as the dominant disease in each segment. Dominant can be defined as optimum number of patients suffered by one disease.

3. Result and Discussion

A total of 14924 patients were attended the GOPD in institute during the period 2015-16 to analyzed. After accessed the database in through 'R' software selected the patients demographic and socioeconomic variables (such as: Age, Sex, Cast, Occupation, Income-Group, Community, Disease, Month of attended) to get the optimum cluster. In Weka software, selected the attributes as Age, Sex, Cast, Occupation, Income-Group, Community, Disease, Month of attended and Month segment to analysis the k-means clustering and to get the results.

3.1 Estimation of optimum clustering

The result of gap statistic is shown in Figure-1. The red dotted line is representing the optimum number of cluster. The optimum estimated number of clusters is 4. Figure-2 show the classification of the clusters against the two principal components (Dim1 and Dim2), cluster 3 & 4 shows well classified but cluster 1&2 shows not well classified because of that some common feature contained both of two clusters.

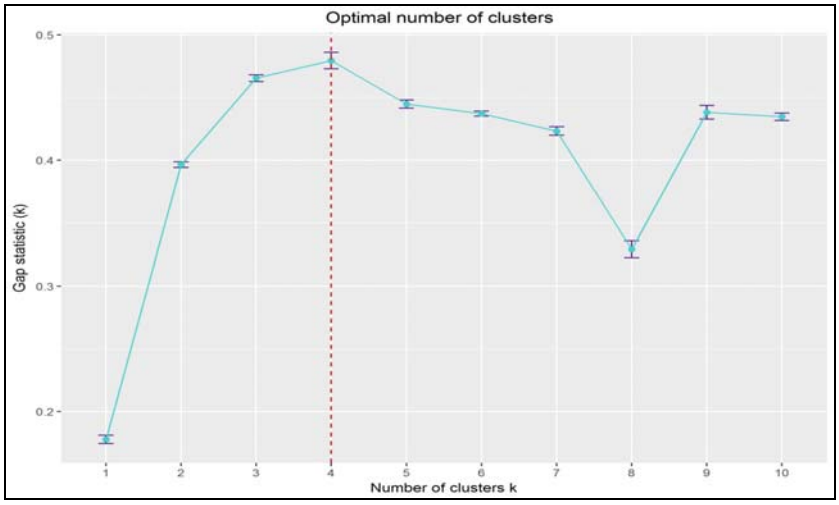

Fig 1: Optimum no. of cluster via gap statistic

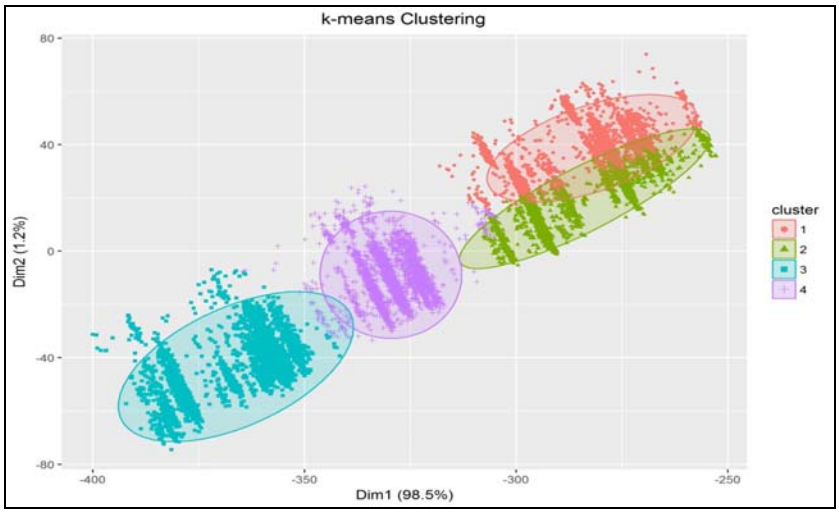

Fig 2: Cluster classification via principal components

3.2 Profiling of Patients: From table-1 shows, that how many patients were attended GOPD in every month segment (three months segment) and from these months segment in which month segment no. of attended patients are optimum. Cluster-1 contain 4229 patients and optimum number of 2465 patients attended GOPD in Oct - Dec month segment. Cluster-2 contains 4787 patients and optimum number of 1713 patients attended GOPD in Jan-Mar month segment. Cluster-3 contains 3299 patients and optimum number of 1111 patients attended GOPD in Apr-Jun month segment. Cluster-4 contains 2609 patients and optimum number of 518 patients attended GOPD in Jul-Sep month segment.

Table 1: Profiling of Attended Patients - Clusters formation

Month Segment	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster of optimum no. of patients for particular month segment
Jan-Mar	588	1713	1265	651	Cluster-2
Apr-Jun	651	1653	1111	1072	Cluster-3
Jul-Sep	525	648	406	518	Cluster-4
Oct-Dec	2465	773	517	368	Cluster-1
Total	4229	4787	3299	2609	

3.3 Disease Pattern Discover via Clustering

From table-2 shows, that the dominant disease (optimum number of patients were attended in GOPD of a particular disease) among all type of disease in particular month segment. Cluster-1 represents the optimum number of patients was suffered from Joint Pain attended in Oct-Dec month segment. Depending upon the Mādda predominates, affecting the joints, Balgham (Phlegm) have the cold and wet properties [8]. Which is responsible for the joint pain according to unani medicine. In this month segment coldness and wetness increased in the environment which enhances the problem of joint pain. That's way the number of joint pain patients increased in this month segment.

Cluster-2 represents the optimum number of patients was suffered from Cough attended in Jan-Mar month segment. Cough is produced due to narrowing of the airways caused by accumulation of secretion and is more prevalent in the persons of Balghami Mizaj (Phlegamatic Temperament) [9]. The derangement of Balgham (Phlegm) is responsible for the cough and Balgham (Phlegm) have cold and wet properties, and in this month segment coldness and wetness increased in the environment which enhances the problem of cough. The increased number of patients is the result of the environmental changes during this month segment.

Cluster-3 represents the optimum numbers of patients were suffered from Piles attended in Apr-Jun month segment. Which is due to the excessive heat in this month segment, excessive heat causes loss of body fluids which causes the constipation mainly and the constipation is the leading cause of piles [10]. Therefore, number of patients are increased in above month segment suffering from piles.

Cluster-4 represents the optimum numbers of patients were suffered from Skin Disease attended in Jul-Sep month segment. In this month segment, temperature, humidity, ultraviolet radiation (UVR), flora and fauna are all change. Patients with fungal infections comprised majority in summer. It is known that warm, humid climates create the environment for the development of fungal infections [11]. During this month segment rain and heat in Bihar, both remains on their peak which is a favourable climatic condition for the skin infections/ailments. Therefore, number of patients are increased in above month segment suffering from skin disorders.

Table 2: Disease Pattern via Clustering

Dominant Disease	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster of optimum no. of patients suffered from particular disease	Month Segment
Joint pain	1069	1053	759	760	Cluster-1	Oct-Dec
Cough	448	636	372	302	Cluster-2	Jan-Mar
Piles	150	166	311	76	Cluster-3	Apr-Jun
Skin Disease	321	280	307	247	Cluster-4	Jul-Sep

Table-3 represents the dominant month segment with dominant disease. The total number of 935 patients were suffered from Joint Pain in Oct-Dec month among of these optimum month was October, 514 patients were suffered from Cough in Jan-Mar month among of these optimum month was January, 210 patients were suffered from Piles in Apr-Jun month among of these optimum month was April, 277 patients were suffered from Skin Disease in Jul-Sep month among of these optimum month was August.

Table 3: Disease Pattern for Dominant Month

Dominant disease	Month segment	Month	No. of patients	Dominant month
Joint Pain	Oct-Dec	Oct.	383	October
		Nov.	223	
		Dec.	329	
Cough	Jan-Mar	Jan.	198	January
		Feb.	163	
		Mar.	153	
Piles	Apr-Jun	Apr.	95	April
		May	90	
		Jun.	25	
Skin Disease	Jul-Sep	Jul.	69	August
		Aug.	122	
		Sep.	86	

4. Conclusion

From this study, to make a profile of patients and discover the disease patterns by cluster analysis. In our country there are so many diseases which are quite impossible to count on fingers even, and we have population in millions. That is the major problem for our government and private institutes to provide the treatment to all in such a huge amount. Usually it is seen that there is always shortage issue of medicines and other resources in institutes to provide treatment to all type of patients. But it will be quite easier to provide treatment to all patients, if the institute will already aware about most spreading disease in a particular months or season. In this regards institute can already arrange all resources in advance and then no patient can be suffer due to the lack of resources and to provide them to batter treatment.

This study provides information about the patient that how many patients were attended the institute, their type of disease and in which month (s) that particular diseased patient attended the most. Through this study to identify disease nature of the patients associated with demographic and socioeconomic characteristics. Such that the institutes come to know the requirements of number of resources for patients of particular disease in that particular month or month segment and will able to fulfil all requirement in advance.

5. References

1. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining, American Association for Artificial Intelligence Menlo Park, CA, USA. 1996, 1-34.
2. McGregor C, Christina C, Andrew J. A process mining driven framework for clinical guideline improvement in critical care. Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). 2012, 765. (http://ceur-ws.org)
3. Ripley B. ODBC Connectivity. Department of Statistics, University of Oxford. 2016. (https://cran.rproject.org/web/packages/RODBC/vignettes/RODBC.pdf)
4. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap statistic. Journal of the Royal Statistical Society B. 2001;

63:411-423.
5. Per Broberg. SAGx: Statistical Analysis of the GeneChip. R package version 1.9.7. 2006. (http://home.swipnet.se/pibroberg/expression_hemsida1.html)
6. Hoyam Omer AA, Saif Eldin FO, Tariq AK. Data Mining of Epidemical Diseases Using Clustering (The Caseof Tropical Disease Institute Omdurman). International Journal of Science and Research (IJSR). 2013; 5(10):340-41.
7. Hartigan JA, Wong MA. A K-means clustering algorithm. Applied Statistics. 1979; 28:100-108.
8. Rabban Tabari, Abu'l Hasan 'Ali ibn Sahl. Firdaws al-Hikma fi'l, Tibb by Hakīm Sayyid, Ashraf Nadvi. Diamond Publication, Lahore, 1992, 818-820.
9. Sehar N, Alam MI, Arfin S, Ahmad T, Ahmad MW, Goswami A. Clinical Study of a Unani Formulation 'Sharbat Zoofa Murakkab' in the Management of Sual Ratab(Productive Cough). Hippocratic Journal of Unani Medicine. 2015; 10(3):1-8.
10. Gami B, Hemorrhoids A. Common Ailment Among Adults, Causes & Treatment: A Reviews. International Journal of Pharmacy and Pharmaceutical Sciences. 2011; 3(5):5-12.
11. Hay RJ, Moore MK. Mycology, in Burns T, Breathnach S, Cox N, Griffiths C, editors. Rook's Textbook of Dermatology, 7th Edition, Oxford, UK, Blackwell Science Ltd. 2004; 31:23.